1 Introduction

1.1 Motivation

Proteins are essential molecules for living organisms. It is approximated that a human body contains between 80,000 and 400,000 such proteins. They are involved in many different functions of the human body including, but not limited to, transporting oxygen in the blood, catalyzing chemical reactions and providing protection against pathogens. Each protein is made up of a unique sequence of amino acids. Various inexpensive and fast experimental methods have been devised over the years that can determine the amino acid sequence from the protein itself. The particular amino acid sequence determines the physiochemical properties of a protein which, in turn, determine how a protein folds and what three-dimensional structure takes. The specific shape of a protein is related to its unique functionality.

A protein domain is a unit of a protein that folds and functions independently from the protein. Figure 1.1 shows an example of the structure of a protein highlighting its three domains. Protein domain boundaries are the boundaries at which a protein domain starts and ends. Identifying these boundaries can provide important information about the overall structure and function of the protein as well as aid in predicting interactions between proteins.

Understanding the structure of proteins is of utmost importance. Designing drugs that can target particular proteins can aid drug discovery, engineering new proteins with desired properties can be used in biotechnology applications and understanding how proteins misfold and aggregate can help develop new treatments for diseases such as Parkinson's disease and Alzheimer. Along with these, there are many other cases in which a protein's structure understanding can be of use.

Experimental methods for determining protein structures exist and can be highly accurate. However, they are associated with many challenges and provide an incentive to explore new methods for the task. A big limitation of experimental methods is how time consuming and expensive they are. Additionally, some proteins are very difficult to study as they are too complex, hard to purify or too unstable. Therefore, new methods for protein structure prediction and understanding that can overcome these limitations are in high demand.

1.2 Aims

In light of the above motivations, we will explore how well a deep neural network can predict protein domain boundaries from the amino acid sequence using different sequence encoding mechanisms. To achieve this we will:

- describe the problem of protein structure prediction from the amino acid sequence using machine learning based approaches.
- collect and process adequate data that can be used to train a deep neural network.
- implement and train a deep neural network with an architecture best suited for the problem domain using different amino acid sequence encoding mechanisms.
- compare the model's performance when trained with different training data.

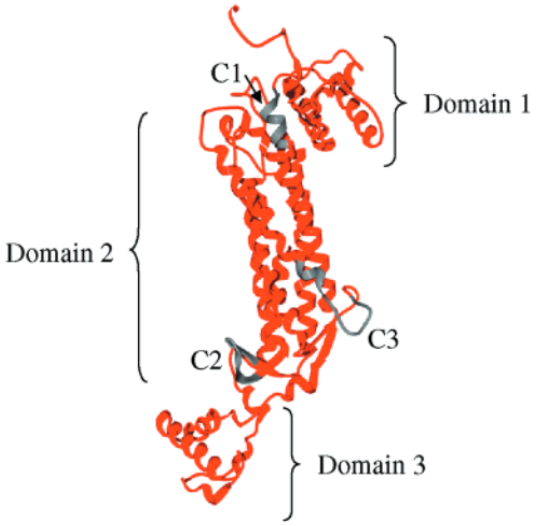

Figure 1.1: *Sampaleanu et al. (2001). A schematic diagram of the three dimensional topology of a protein indicating its three structural domains.*

- evaluate and analyse the models' results and compare them with other similar machine learning methods that exist in literature.
- discuss the limitations of our methods and suggest possible improvements that could have been adapted in our solution.

(should I add a motivation for doing this here or later?)

Furthermore, we want to evaluate evaluate how AlphaFold, a novel algorithm designed for predicting the entire 3D structure of a protein from the amino acid sequence, performs in the protein domain boundary prediction task. This task is not trivial and requires adequate attention. We aim to do the following:

- describe the problem of evaluating AlphaFold for the domain boundary prediction task.
- collect appropriate AlphaFold-predicted 3D structures and translate them to AlphaFold-predicted domain boundaries.
- evaluate the results and compare them with results obtained from methods devised for explicitly predicting domain boundaries.
- discuss our findings and the limitations of our process.

2 Background

2.1 Deep Learning

The success of Deep Learning in different problem domains has drawn the attention of researchers

2.2 Protein Representations

ESM, CARP etc

2.3 Predicting protein structures

2.3.1 Domain Boundary Prediction

Wang et al. (2022) explored and evaluated the performance of a novel algorithm , namely Res-Dom, for protein domain boundary prediction. They extract four features from the amino acid sequence: solvent accessibility, secondary structure, a Hidden Markov model profile and the sequence's embedded features from a pre-trained protein language model. Then, a Residual Neural Network, followed by a bi-directional long short-term memory (BLSTM), take these four features as input and output the predicted domain boundaries. Their evaluation shows that this method can achieve highly accurate results. The limitation of this method is that it depends on other methods for extracting features from the amino acid sequence. For example, the work of Cheng et al. (2005) used to extract the secondary structure and solvent accessibility can only work on proteins with a sequence length with less than 1500 residues. In turn, this limitation is also inherited by Res-Dom for predicting protein domain boundaries.

Shi et al. (2019) present a method for predicting domain boundaries and argue that local and non-local interactions between residues provide useful information for the overall structure of the protein. This is taken into account and a method that captures both local and non-local interactions has been devised using a multi-channel Convolutional Neural Network (CNN) and a stacked bidirectional Gated Recurrent Unit (BGRU) In doing so, the work of Cheng et al. (2005) is again present which also introduces the limitation on the number of residues for a sequence.

Jiang et al. (2019) have followed an *ab initio* approach without using third party methods which can introduce limitations such as the one seen in Res-Dom or DNN-Dom. This method only depends on the amino acid sequence of proteins which is used to train a stacked BLSTM. Each amino acid sequence was encoded using 5 numerical descriptors collected from a comprehensive list compiled from public databases. Their results demonstrate that, although this method is faster than template-based methods and has no dependencies, the accuracy is lower.

Cretin et al. (2022) have taken a structure-based approach.

2.3.2 Domain Prediction

The work of Zheng et al. (2020) offers its own contribution in protein structure prediction by predicting the domain each residue belongs to which provides more information about the structure of the protein than the domain boundaries do.

2.3.3 3D Structure Prediction

AlphaFold 2 (AF2), the work of Jumper et al. (2021), is the most successful and accurate method to date for predicting the entire 3D structure of a protein according to the 14th Crticial Assessment of Structural Prediction (CASP14) competition. This method is based on deep learning and involves complex novelties. The network consists of two main components. the first utilizes the work of Vaswani et al. (2017) to generate contextual representations from the input data: a contact map, which captures which residues interact with each other, and representation of the chain's homologous sequences. The second is responsible for generating the 3D structure of the protein utilizing the generated representations, and by utilizing an ensemble of multiple prediction models to improve its accuracy. Even though AF2 demonstrates promising capabilities, its practical use is under exploration, including protein domain boundary assignment.

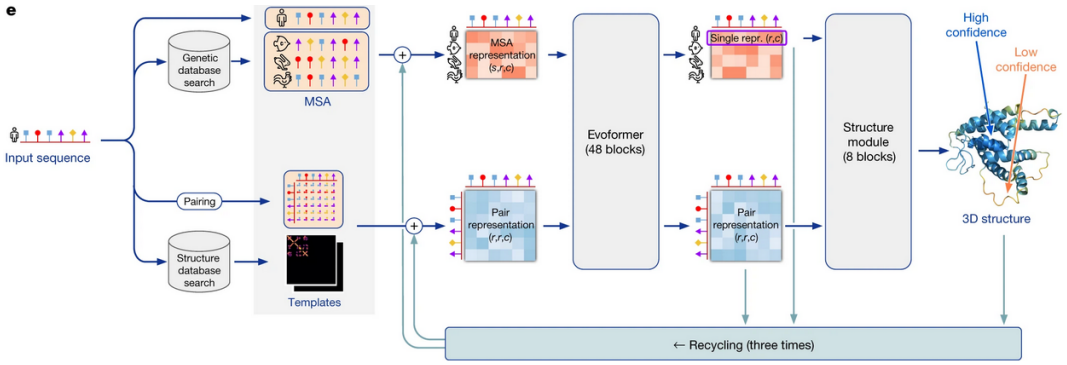

Figure 2.1: AlphaFold

2.4 Deep Neural Networks

2.4.1 Recurrent Neural Networks

Min et al. (2016) have shown